Weekly Report

Lu Junhua

2015 年 8 月 30 日

This week, I spend most of my time in Gongan, from 10am to 6pm, from Monday to Friday. We(student from Prof. Pan and I) completed the plan made by Yinghua He. What I did:

- revise the function of getting first-crime data attribute.Since the crime date store in the Solr is not sorted by real crime time, but the info start-using time, so we have to re-sort the date. However, the record is read one by one, so each time a new date is read, we compare it to the former one, and keep the one of earlier time. And thus, we get the earlist crime time. After revising, the efficient of drawing data fall down a little.(Since Ke, Feng didn't leave any log last week, I don't know clearly about the efficiency.)

- Extract data again. Since there existed many data missing in the system of Gongan and Feng is too stubborn to listen to our advice, we wasted much time on drawing data. This week, I re-extract them and check whether the data had any mistakes. Now I believe most remaining errors are caused by the system of gongan, not ours.

- We set about to run the estimation code on Wednesday. I introduced the data, the data structure to the new member Zhu Dandan, and study the code given by Prof.He. Zhu is more familiar with STATA than I, and I found some little bugs in the code. Most of my job is checking the consistency of contextual codes, and got to know some syntax of STATA. And due to the low version of STATA, we change the STATA to STATA13 multiple-processor version, which provide larger ram and more cores for processing data. It increase our efficiency a lot.

- We phoned with Prof.He two times a day, through mobile phone and Skype,on 3pm and 10.30 pm, for raising and solving problems, read and revise and interpret stata codes. At any other time, we chat on Wechat.

The result can be referred to on Saturdays email. Here is some screenshots.
Besides this, I read the paper Interactive Visualizations for Deep Learning and Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. The former one is a workshop paper, and the later is literature review. The review is writ six years ago, which proposed many problems,ideas and suggestions for both data mining and info-vis. Many problems remains to be solved, while others are solved by now. This paper summarized many papers at that time, and worth using for reference.

摘要

本文简要总结了犯罪预测项目的初步结果。利用近 15 万人 2013、2014 年的数据，我们研究了一个 logit 离散选择模型；模型的预测准确率在两个方面得到很好的体现：一、在“估计样本（训练样本）”中，如果只用 2013 年的数据来预测 2014 年的犯罪情况，在模型判定为犯罪概率高的人士中，43% 在 2014 年最终犯罪；二、我们用“预测样本（未用于估计的数据）”来模拟这个模型的具体应用——假设现在的时间是 2014 年 12 月 31 日，我们用 2013、2014 两年的数据来预测人们未来是否犯罪；模型找出的犯罪概率高的人士之中 38%在 2015 年 1-6 月成为罪犯。

根据这些成果，如果现在把这个模型应用于整个杭州，在一定假设条件下，我们推算：模型可以找出 30,090 名犯罪概率高的人士，其中有 **11,434 人**（38%）**在 2015 年 9 月至 2016 年 2 月会真的犯罪**。

同时，我们可以调整阈值来使得需要关注的总人数减少；比如，模型可以找出 2,760 犯罪概率高的人士，其中有 1579 人（42%）在 2015 年 9 月至 2016 年 2 月会真的犯罪。

表 1：估计样本中的犯罪预测

2013 年平均每月犯罪概率阈值	预测可能犯罪的人总数	2014 年真实犯罪情况	
		犯罪	未犯罪
0.01	1480	639	841
		(43%)	(57%)
0.045	161	65	82
		(40%)	(60%)

表 2：预测样本中的犯罪预测

2014 年平均每月犯罪概率阈值	预测可能犯罪的人总数	2015 年 1-6 月真实犯罪情况	
		犯罪	未犯罪
0.01	1003	385	618
		(38%)	(62%)
0.045	92	39	53
		(42%)	(58)

Next week, I may still assist Zhu in processing data, if necessary. And I will read at least 2 more papers and will organize a explicit frame of predictive visual analytics.